EXECUTIVE SUMMARY

Business Problem

Hospitality industry thrives on customer feedback and customer online reviews makes a huge impact on the brand value of a service offered by a hospitality service provider. In the recent era of globalization, TripAdvisor is the popular website that most customers use for booking hotels online and also provide their reviews. It is imperative for the hotel service providers to understand and analyze the customer sentiment about a service in the hotel and improve their services to increase their occupancy and withstand the competition. In addition, the increasing amount of reviews pose a challenge on hotel management on what reviews to consider that are worth improving their reputation.

As a result, Hospitality service providers are in search of finding answers to some of the key questions impacting their business as listed below.

• Strategy to run chain of hotels across many cities?

• What drives customer to reserve a hotel booking?

• What turns customer away from choosing a hotel?

• How do I balance occupancy and price to maximize revenue?

• Which segment do I operate in?

• Which segment should I compete with?

• How is my performance over a period of time?

Primary objective is to do text mining on TripAdvisor customer Reviews and create business value for hotel management. Our Project aims to analyze and provide useful insights for hotels at different levels and answer the above questions.

 Prominent factors for Hotels located in different demographics

 In-depth Analysis of hotels in a particular city

 Analysis for individual hotel

These insights will help them improve and further innovate the services offered.

Text mining:

Preprocessing of textual data:

Corpus: A Corpus is a large collection of documents which are related to a similar organization or a similar author. These corpuses are used to perform the basic statistical operations such as frequency counts of terms, retrieving the instances of a particular occurrence. This is used in the Natural language processing.

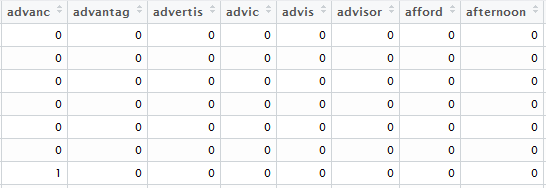
There are few steps followed by us to process the Corpus to improve the quality of textual data we have.

1. Generating the corpus using the textual data we have in the form of customer reviews for the hotels.
2. Converting all the textual data to lower case which helps the document to be plain and useful in searching for the text.
3. I am going to convert this to plain text document to make the corpus stable.
4. As we are using hotel reviews there is a high possibility that customers use numbers such as bill amount, room number in the reviews, this might be the noise in our data, to reduce this I am removing all the numbers in the reviews.
5. In the next step I have removed stops words, typically stop words are the words which are used quite often and they will not be useful to classify the document. I have dealt with these by removing both English stop words and custom initialized stop words from the data.
6. In the next step we have performed stemming on the words in corpus, in other words we can say it as grouping of similar words to a single root. Let us take three words ‘similar’,’similarly’,’similarity’, if you observe all these words have same meaning but they are used in different context. Usually NLP tools take these words as separate words, which effects our word frequencies. To increase the weights and classification ability of these words we can use a single word to determine all three, such as ‘similar’. This will not only increase the word count but also provides better classification rate with this word.

Document term matrix and frequencies:

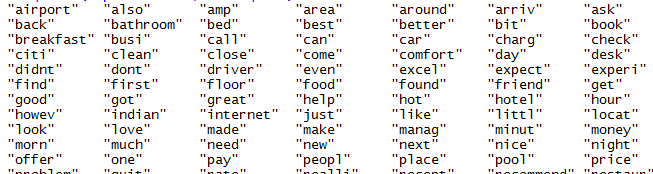
A document term matrix is a frequency of terms matrix that is generated using the collection of documents called Corpus in text mining. This has been the primary basis for the natural language processing(NLP). Document term matrix is also termed as a sparse matrix which has 1’s and 0’s for all the terms as columns in it, ‘1’ indicates that a word is present in the document and ‘0’ indicates that a word is not present. Pic 1.1 shows a typical document term matrix(DTM), in our case we used all the reviews by customers to the hotels across 8 different cities in the world to generate the matrix. DTM has been the founding step to perform multivariate analysis such as clustering, Linear Discriminant Analysis (LDA) on textual data, this will be explained in the coming sections. We have selected the words which are at least repeated in 1% of the data. This helps us to capture better classifiers and we can reduce the unwanted terms from the data. This process is called removing sparse terms.

1.1 Typical document Term Matrix

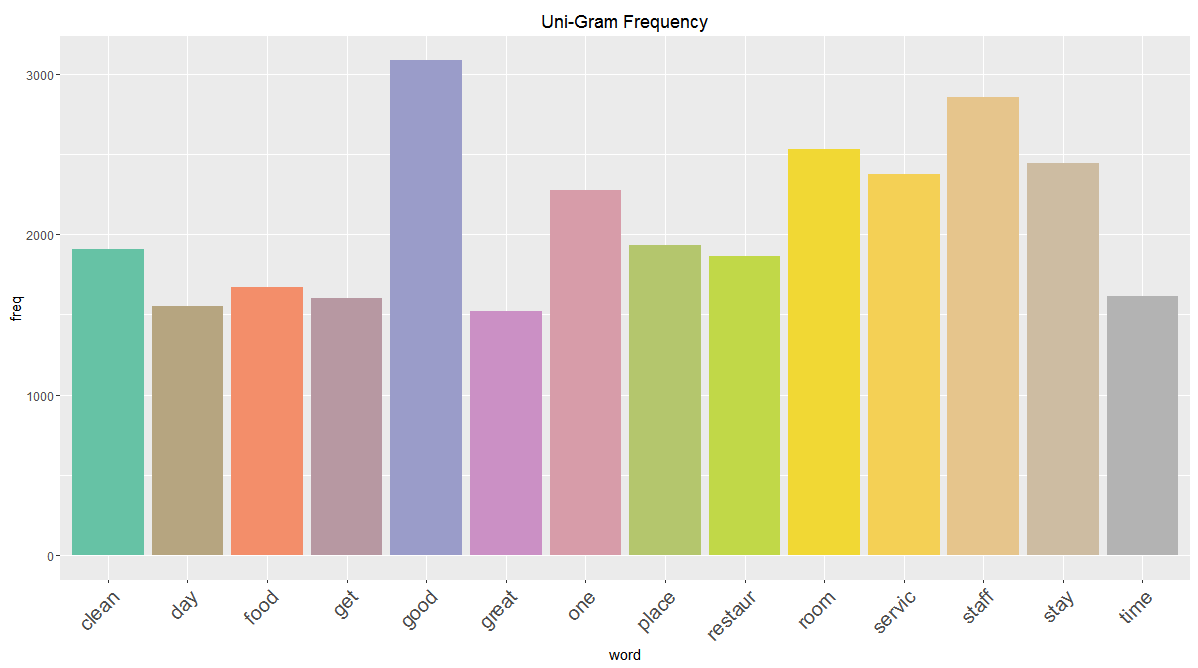


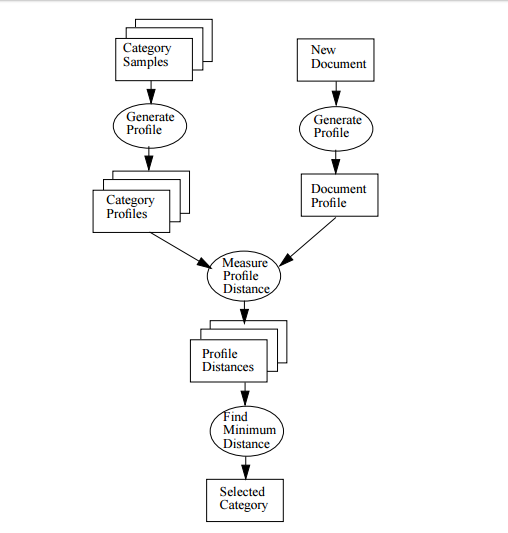
This helps us to determine what words are prominent in reviews, probably most weighted terms or words used the most number of times in the user reviews. A better example in the hotel reviews is shown in the below diagram, these are the words which are repeated at least 500 times in the reviews for Delhi hotels. If we observe words like “area, clean, breakfast, charge, comfort, great, friend’ shows positive attitude towards the hotel, where as words like “hot, didn’t, better, little’ explains negative impact. No wonder a hotel which is liked by someone is not good for another person.

Words used at least for 500 times in all Delhi reviews.



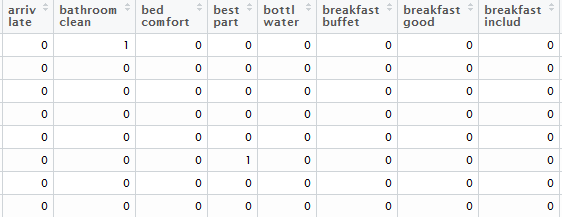
The same is explained using the UNI-gram frequencies, where I have calculated the number of times a word is repeated in the reviews, the words in the picture below are at least used more than 1500 times by the customers in their reviews. I have used 4300 Delhi customer reviews to analyze the patterns. I observed most of the time people used words like “clean, good, food, service, staff, stay”, these are the basic things everyone will look when they want to book a hotel. For a beginner in hotel industry this information is crucial to come up with the quality of services they want to provide to the customers.

 However, by using single words we will not be able to tell whether these are appreciating the services offered or complaining about services offered, to study this in detail we have used a theory called N-Gram-Based Text Categorization, where N is combination of words required to be formed. Typical data flow for the N-gram text categorization is shown in the below picture,

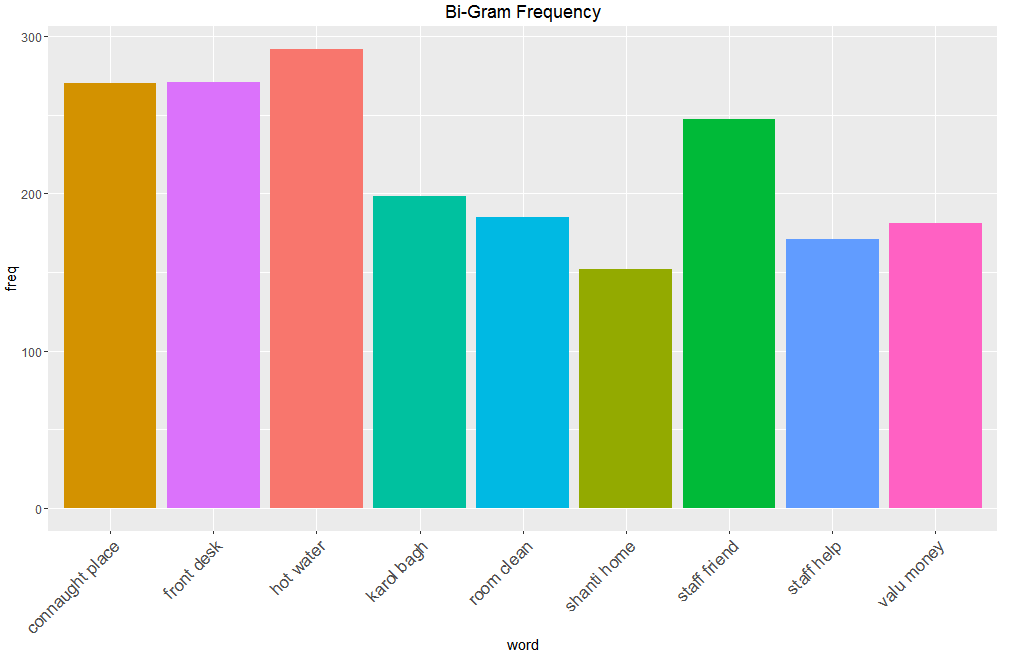


Typically, after cleaning the data in the corpus with the above stated text pre-processing steps, we create a DTM using N-grams where N (Integer). This iteratively computes the distances between words among the documents and selects the word combinations the customer used the most based the N.

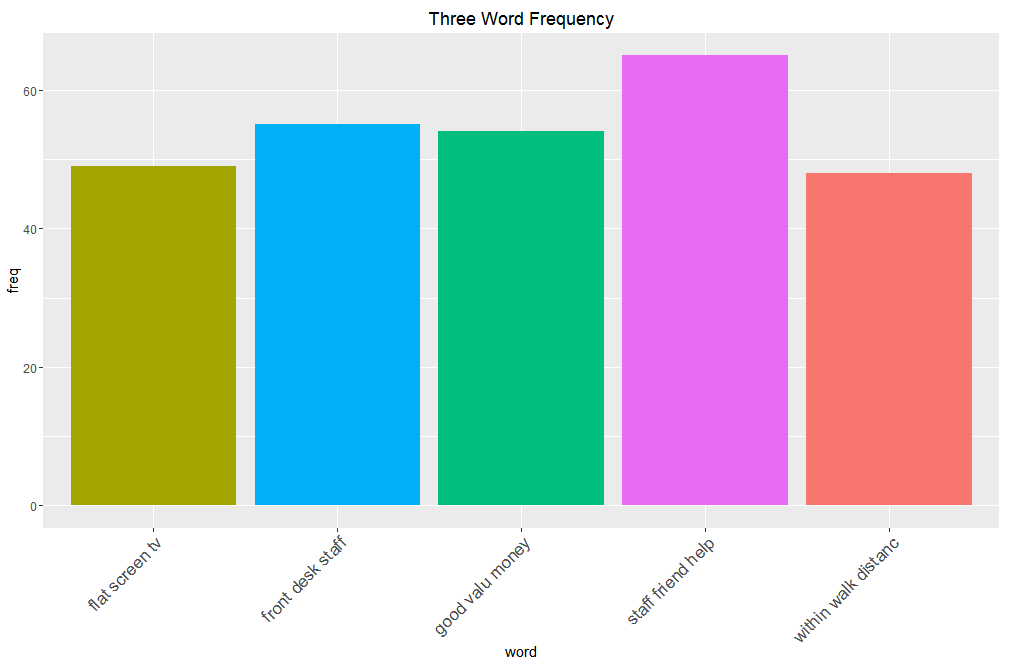
I have computed the N-gram frequencies on Delhi reviews data with N values (1,2). Picture below shows two-word combination document term matrix that is generated. This has more insights because two words tend to give a meaningful sentence, for example if we take breakfast good variable, when this is 1 for a document, the customer is saying they liked breakfast in here. This type of information if useful for the future customers.



A more detailed visual is shown below, these are the word combinations which N-gram computed and has a frequency of more than 150 times in the documents, Words like ‘carol bagh, shanti home’ are place names, this helps us removing these words as stop words from the data. More over words like ‘room clean, staff help, valu money’ shows positive attitude towards the hotel.



We have also used TRI-gram computations to find more trends among the user reviews, these categories helped us to initially categorize among hotels, if we can observe few hotels review were having comments “good value money, staff friend help, within walking distance”, these are positive words which are helps customers know more about the hotels.



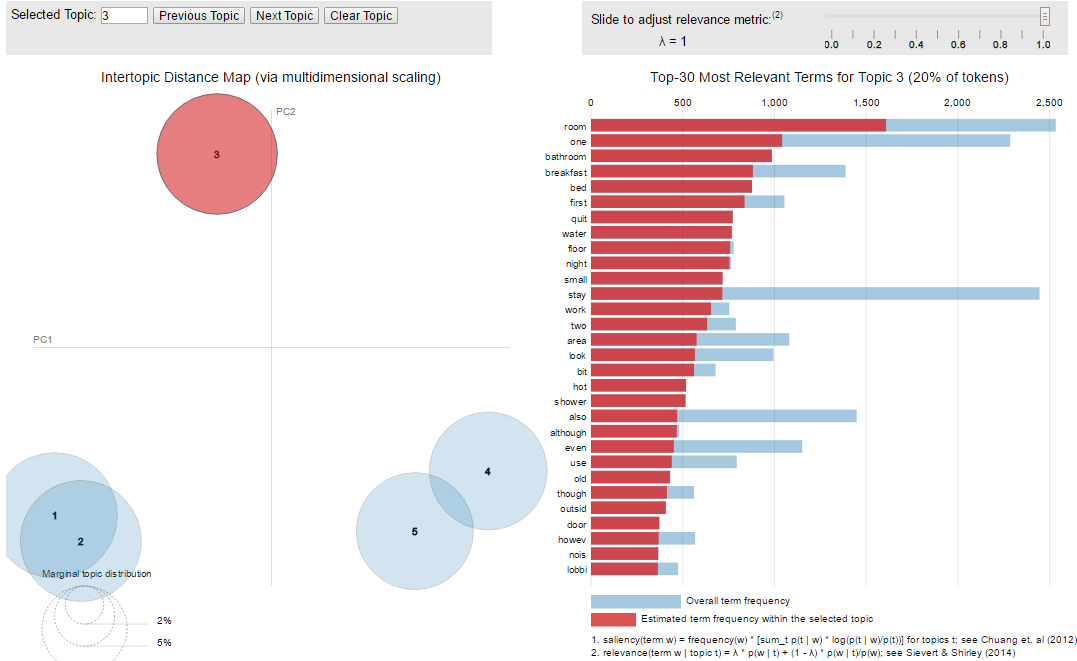
Topic Models

LDA:

Latent Dirchelet Allocation is the process of topic modelling technique we used to segregate between the Reviews, with help of Gibbs sampling we created five different topics for the hotels in city Delhi. I have incorporated a folder of the web application, please use it in a Mozilla browser to get more insights. Below are the text topics we generated for the hotels in Delhi, From the LDA visualization and the normal tabular forms.

We can clearly see that distances between the topics 3 and rest is very large, as we assumed to our surprise The topic 3 has few negative words such as ‘noise, tired, lack, dirty’ and many more which you can effectively observe using the web application we provided. Using this data combining with individual hotel rating we determined that reviews in the topic three are more towards negative sentiment. Moreover, Topics 4 and 5 have words such as ‘service, good, clean ‘stating positive attitude towards the hotels.

Latent Dirchelet Allocation of Topic Models Visulization.



References:

1. http://odur.let.rug.nl/~vannoord/TextCat/textcat.pdf

2. LDA Visual from CRAN.